Models for Queue Length in Clocked Queueing Networks

by

Ora E. Percus
J. K. Percus

Ultracomputer Note #153
January, 1989

This work was supported in part by the Applied Mathematical Sciences Program of the US Department of Energy under contract DE-FG02-88ER25052

1. Introduction

Digital computers in practice are synchronous machines running on a clock cycle, a mode of operation which is mandatory when parallel operations are to be performed. The information that is routed from node to node through the internal networks of an efficiently designed machine will not always be transmitted at once through a node, but may have to wait, effectively in a queue. Although the analysis of networks with queues has an ancient and honorable history, this analysis has become routine only for stationary operation in continuous time. Transient problems or, much more importantly from our present point of view, discrete clock-regulated transmissions, have been considered in the main from the direction of general existence theorems, or for highly oversimplified models.

We would like in this paper to make a start towards systematic analysis of discrete-time networks with queues. In conformity with several current network realizations [1], the unit element of the networks we consider will be the 2×2 buffered switch, which we can regard as a system of two queues working in parallel, each one with a deterministic server of unit serving time. A message

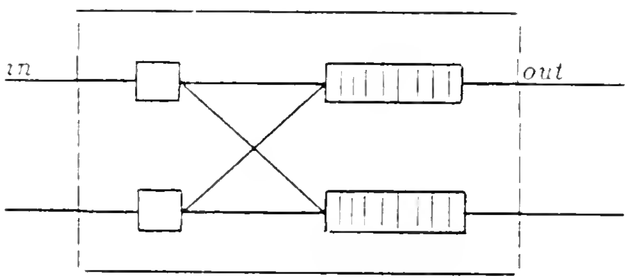

Fig. 1. 2×2 buffered switch

entering either of the two inputs of the switch goes with probability 1/2 to either of the two output queues of the switch. Any node in this network will see a time series, each element of which is the presence or absence of an information packet to be transmitted to the next node. Thus we are observing the transformation of discrete time series by the basic processes of passage through a queue and of mixing.

Continuous time stationary queuing systems in such as Jackson networks [5] are characterized as Poisson processes, only one parameter being necessary to describe which member of the class is

under consideration, and with the guarantee that elementary transformation processes do not take one out of this class. The situation is very different with discrete clock-regulated networks, even for stationary operation, which will be our principal concern. For example, it has been shown that if independent uncorrelated Bernoulli streams of 0's and 1's (packet absent or present) enter the input of the switch of Fig. 1, then each output is instead a highly correlated multi-parameter renewal process [4], and subsequent mixing at the next input complicates this further. Ultimately, one would like to work with the full class of time series encountered and completely define its network transformations. A more modest approach (other quite effective empirical techniques exist in the literature [3]), which we will adopt in this paper, is that of modeling the time series by classes which can be readily handled and upon which any time series encountered can be "projected" (in the Galerkin sense [2]). As the parameter space of the model class is enlarged, higher and higher accuracy will then be obtained.

In outline, we first present the basic single queue relations and carry out elementary manipulations to obtain "ball-park" bounds on mean queue length. We then show how to transform transition probability information from the output of one switch to the input of the next one. With this as background, we carry out the Markov time series input model to predict stage to stage mean queue length evolution, as well as asymptotic values. These estimates turn out to be uniformly lower bounds. We then proceed to a renewal process input time series model, applying it only to the "second stage" problem, where it yields uniformly an upper bound on mean queue length. We conclude with suggestions for feasible generalizations.

2. Preliminary Estimates

Suppose the inputs to the switch of Fig. 1 at time n are denoted by I_n and I'_n, each taking the value 0 or 1. If we confine our attention to the upper channel, then according to the rule for mixing,

$$a_{n+1} = \gamma_n I_n + \gamma'_n I'_n \tag{2.1}$$

will arrive at the head of the upper queue, where

$$\gamma_n = \begin{cases} 0 \\ & \text{with probability } \frac{1}{2} \\ 1 \end{cases} \tag{2.2}$$

and similarly for γ'_n; the streams $\{\gamma_n\}$ and $\{\gamma'_n\}$ are i.i.d. random variables. Thus, if the queue had length S_n to start, it would now contain

$$C_{n+1} = S_n + a_{n+1} \tag{2.3}$$

One packet is served – transmitted – as long as $C_{n+1} > 0$, but none of course if $C_{n+1} = 0$. Thus at the end of the clock cycle,

$$\begin{aligned} S_{n+1} &= S_n + a_{n+1} - 1 + \delta_{n+1} \\ \text{where } \delta_{n+1} &= \begin{cases} 0 & & C_{n+1} > 0 \\ & \text{if} & \\ 1 & & C_{n+1} = 0\,, \end{cases} \end{aligned} \tag{2.4}$$

or in compact form

$$S_{n+1} = \max\{S_n + a_{n+1} - 1.0\}. \tag{2.5}$$

Furthermore, the output is given by

$$O_{n+1} = 1 - \delta_{n+1}. \tag{2.6}$$

serving as input for the next stage.

Evidently, the above can be taken over unchanged to the operation of the lower channel and associated queue (γ and γ' are to be replaced by their complements). Although the outputs of the two channels are clearly correlated, matters may be arranged, e.g. in the Ω–network, so that they

and their progeny never meet again as joint inputs, and so this correlation will not be investigated. It is in fact a problem which is analytically quite involved.

In steady state, or stationary, operation the rate of information flow in each branch is constant, with expectation, say $E(I_n) = \rho < 1$. Since $I_n = 0$ or 1, this implies that

$$
\begin{aligned}
P(I_n = 0) &= 1 - \rho \\
P(I_n = 1) &= \rho,
\end{aligned}
\tag{2.7}
$$

and of course O_{n+1}, by conservation, has the same property. On the othe hand, we see from definition (2.1) that

$$
\begin{aligned}
f_2 &\equiv P(a_{n+1} = 2) = (\rho/2)^2 \\
f_1 &\equiv P(a_{n+1} = 1) = \rho(1 - \frac{1}{2}\rho) \\
f_0 &\equiv P(a_{n+1} = 0) = (1 - \frac{1}{2}\rho)^2.
\end{aligned}
\tag{2.8}
$$

But (2.7) and (2.8) do not define either time series, which require multiple autocorrelations for their complete characterization. A question of obvious interest is how much information is available on the basis of (2.8) alone. In particular, we will concentrate on the mean queue length $E(S_n)$, a crucial parameter for designing switch capacity. A weak bound is in fact easily obtained.

Consider the basic relation (2.4). Since

$$
\begin{aligned}
S_n\delta_{n+1} = 0, \quad a_{n+1}\delta_{n+1} &= 0 \\
\delta_{n+1}^2 &= \delta_{n+1},
\end{aligned}
\tag{2.9}
$$

multiplying (2.4) by δ_{n+1} tells us that

$$
S_{n+1}\delta_{n+1} = 0, \tag{2.10}
$$

a fact which is obvious without any algebra. Now write (2.4) as $S_{n+1} - \delta_{n+1} = S_n + a_{n+1} - 1$ and square both sides:

$$
S_{n+1}^2 + \delta_{n+1} = S_n^2 + 2S_n(a_{n+1} - 1) + (a_{n+1} - 1)^2. \tag{2.11}
$$

In steady state, $E(S_{n+1}^2) = E(S_n^2)$, so that (2.11) implies

$$
E(S_n) = E(a_{n+1}S_n) + \frac{1}{2}E(a_{n+1} - 1)^2 - \frac{1}{2}E(\delta_{n+1}). \tag{2.12}
$$

According to (2.4), $E(\delta_{n+1}) = E(1 - a_{n+1})$. Computing expectations via (2.8), we conclude that

$$
\begin{aligned}
E(S_n) &= E(a_{n+1}S_n) + \frac{1}{2}E(a_{n+1}(a_{n+1} - 1)) \\
&= E(a_{n+1}S_n) + \rho^2/4.
\end{aligned} \tag{2.13}
$$

As an immediate consequence,

$$
E(S_n) > \rho^2/4. \tag{2.14}
$$

which turns out to be a very weak bound indeed unless ρ is quite small.

For more incisive information, the specific properties of the time series $\{a_n\}$ must surely be taken into account. For a start along this direction, let us iterate (2.4) to give

$$
S_n = S_0 + \sum_{i=0}^{n-1}(a_{n-i} - 1 + \delta_{n-i}) \tag{2.15}
$$

(we imagine that the process starts at $n = -\infty$) and then insert into (2.13):

$$
E(S_n) - E(a_{n+1}S_0) = \rho^2/4 + \sum_{i=0}^{n-1} E(a_{n+1}(a_{n-i} - 1 + \delta_{n-i})). \tag{2.16}
$$

The stationary character of the process allows us to shift indices by $n - i$ and hence write instead

$$
E(S_n) - E(a_{n+1}S_0) = \rho^2/4 + \sum_{i=1}^{n} E(a_i(a_0 - 1 + \delta_0)) \tag{2.17}
$$

Let us take the limit $n \to \infty$. Asuming that the correlation range of $\{a_n\}$ is not infinite, and noting that S_0 is an explicit function of $a_{-1}, a_{-2}, \ldots$, then S_0 and a_n become independent for large n. Since $E(S_n) = E(S)$ does not depend on n and $E(a_{n+1}) = \rho$, the left hand side of (2.17) becomes $(1 - \rho)E(S)$. On the other hand, since $E(a_i\delta_0) = E(a_i|\delta_0 = 1)P(\delta_0 = 1) = (1 - \rho)E(a_i|\delta_0 = 1)$, we can write

$$
E(a_i(a_0 - 1 + \delta_0)) = [E(a_i a_0) - \rho^2] + (1 - \rho)[E(a_i \mid \delta_0 = 1) - \rho]. \tag{2.18}
$$

The virtue of (2.18) is that $E(a_i a_0) \to E(a_i)E(a_0) = \rho^2$ and $E(a_i \mid \delta_0 = 1) \to E(a_i) = \rho$ as $i \to \infty$, so that both contributions to (2.18) converge to 0 as $i \to \infty$. If the convergence is rapid enough, we may take the limit $n \to \infty$ in (2.17) to arrive at

$$
\begin{aligned}
(1 - \rho)E(S) &= \rho^2/4 + \sum_{i=1}^{\infty}(E(a_i a_0) - \rho^2) \\
&\quad + (1 - \rho)\sum_{i=1}^{\infty}(E(a_i \mid \delta_0 = 1) - \rho).
\end{aligned} \tag{2.19}
$$

the infinite series representing the memory of past queue length.

Unfortunately, (2.19) as it stands makes use not only of the presumably known time series $\{a_n\}$, but also of the specific process in the form of the condition $\delta_0 = 1$. It is hence directly useful only for sufficiently restricted time series that the condition becomes innocuous, and this is precisely what we will take advantage of in the sequel.

3. Output-Input Transformation

We are now prepared to start a detailed analysis of the time series transformation represented by Fig. 1. Of the two components, collection of inputs (2.1), and passage through the queue, the former is far simpler to analyze. Since I_n and I'_n represent outputs of previous switches, this is basically an input-output transformation, and is most readily handled by the use of generating functions.

Suppose that r times $n_1, n_2, \ldots, n_r$ are selected, and that

$$\begin{aligned} G_R(z_{n_1}, \ldots, z_{n_r}) &\equiv \sum_{\{\nu_i = 0,1\}} P(I_{n_1} = \nu_1, \ldots I_{n_r} = \nu_r) z_{n_1}^{\nu_1} \ldots z_{n_r}^{\nu_r} \\ &= E(z_{n_1}^{I_{n_1}} \ldots z_{n_r}^{I_{n_r}}); \end{aligned} \tag{3.1}$$

we assume the same statistics for $\{I'_n\}$. Furthermore, define

$$\begin{aligned} F(z_{n_1}, \ldots z_{n_r}) &\equiv \sum_{\{\alpha_i = 0,1,2\}} P(a_{n_1+1} = \alpha_1, \ldots a_{n_r+1} = \alpha_r) z_{n_1}^{\alpha_1} \ldots z_{n_r}^{\alpha_r} \\ &= E(z_{n_1}^{\gamma_1 I_{n_1} + \gamma'_1 I'_{n_1}} \ldots z_{n_r}^{\gamma_r I_{n_r} + \gamma'_r I'_{n_r}}) \end{aligned} \tag{3.2}$$

But the $\{\gamma_i, \gamma'_i\}$ are i.i.d. Bernoulli with probability $\frac{1}{2}$, and the sequences $\{I_n\}$, $\{I'_n\}$ are independent of each other. Since

$$E(z^{\gamma I} \mid I = \nu) = \frac{1}{2}(1 + z^{\nu}), \tag{3.3}$$

we see at once that

$$F(z_{n_1}, \ldots z_{n_r}) = [E(\frac{1}{2}(1 + z_{n_1}^{I_{n_1}}) \ldots \frac{1}{2}(1 + z_{n_r}^{I_{n_r}})]^2. \tag{3.4}$$

Expanding out the argument of the expectation, we conclude that

$$F(z_{n_1}, \ldots, z_{n_r}) = (\frac{1}{2^r} \sum_{\Lambda \in R} G_\Lambda(z_{n_1}, \ldots, z_{n_r}))^2 \tag{3.5}$$

where $R = \{n_1, \ldots, n_r\}$ and G_Λ is defined as having arguments restricted to those with indices in Λ.

For our present purposes, it will be sufficient to consider two adjacent times, $R = (0, 1)$, so that – dropping superfluous indices – (3.5) reduces simply to

$$F(x, y) = \frac{1}{16}(1 + G(x) + G(y) + G(x, y))^2 \tag{3.6}$$

where $x = z_1$, $y = z_0$

According to (2.7), we have

$$G(x) = 1 + (x-1)\rho. \tag{3.7}$$

As for $G(x,y)$, we must adopt a notation for the joint distribution of I_0 and I_1. It will be convenient to write

$$\begin{aligned} P(I_1 = 1 \mid I_0 = 0) &= \rho_0 \\ P(I_1 = 1 \mid I_0 = 1) &= \rho_1, \end{aligned} \tag{3.8}$$

so that ρ_0 and ρ_1 would both reduce to ρ for independent sequences of inputs. Of course, it follows that

$$\begin{aligned} P(I_1 = 0 \mid I_0 = 0) &= 1 - \rho_0 \\ P(I_1 = 0 \mid I_0 = 1) &= 1 - \rho_1, \end{aligned} \tag{3.9}$$

and hence that

$$\begin{aligned} G(x,y) = (1-\rho_0)(1-\rho) + \rho_0(1-\rho)x \\ + (1-\rho_1)\rho y + \rho\,\rho_1 xy. \end{aligned} \tag{3.10}$$

But ρ_0, ρ_1, and ρ are related. For example, we may eliminate ρ_0 by noting that

$$\rho = P(I_1 = 1) = \rho_0(1-\rho) + \rho_1\rho, \tag{3.11}$$

so that

$$\rho_0 = \frac{\rho}{1-\rho}(1-\rho_1). \tag{3.12}$$

To reduce notational complexity it will now be convenient to introduce complementary probabilities

$$p = \frac{1}{2}\rho, \quad q = 1 - \frac{1}{2}\rho, \tag{3.13}$$

as well as a normalized departure from independence,

$$r = \frac{\rho_1 - \rho}{2 - \rho}, \tag{3.14}$$

whereupon

$$\begin{aligned} \rho_1 &= 2(p + qr) \\ \rho_0 &= \frac{2p}{q-p}(q - p - 2qr). \end{aligned} \tag{3.15}$$

Eq. (3.10) then takes on the more symmetric form

$$G(x,y) = (q-p)^2 + 4pqr + 2p(q-p-2qr)(x+y) + 4p(p+qr)xy, \tag{3.16}$$

in terms of which (3.6) becomes

$$F(x,y) = [q(q+pr) + pq(1-r)(x+y) + p(p+qr)xy]^2. \tag{3.17}$$

To find the joint distribution $P(a_1 = k,\ a_0 = k')$, one need only take the appropriate coefficients in (3.17), via the definition (3.3). The transition probabilities

$$\begin{aligned} P_{kk'} &= P(a_1 = k \mid a_0 = k') \\ &= P(a_1 = k,\ a_0 = k')/f_{k'} \end{aligned} \tag{3.18}$$

will be of direct interest ($f_{k'}$ is taken from (2.8)) and are presented in Table 1.

$k \backslash k'$	0	1	2
0	$(q+pr)^2$	$q(q+pr)(1-r)$	$q^2(1-r)^2$
1	$2p(q+pr)(1-r)$	$2pq(1-r)^2+r$	$2q(p+qr)(1-r)$
2	$p^2(1-r)^2$	$p(p+qr)(1-r)$	$(p+qr)^2$

Table 1. The transition matrix $P_{kk'}$.

4. A Markov Model

In analyzing the switching networks under consideration, we are faced with two related problems. First is the transformation of time series as one goes from input to input, and second is the statistics of the queue length at a given switch. We have focussed thus far on the basic parameters of the time series. $f_k = P(a_n = k)$ and $P_{jk} = P(a_{n+1} = j \mid a_n = k)$. Our intention now is to restrict ourselves to situations for which such information is sufficient. Presumably this will be the case if $\{a_n\}$ is modeled to lie in the class of Markov time series, which we now assume.

The picture we have in mind is that of a *network* of switches. This means that at a given switch

(4.1) $$I_n = \overline{O}_n, \quad I_n' = \overline{O}_n'$$

where $\overline{O}_n$ and $\overline{O}_n'$ are outputs from previous switches. We have seen in Sec. 3 that the parameters

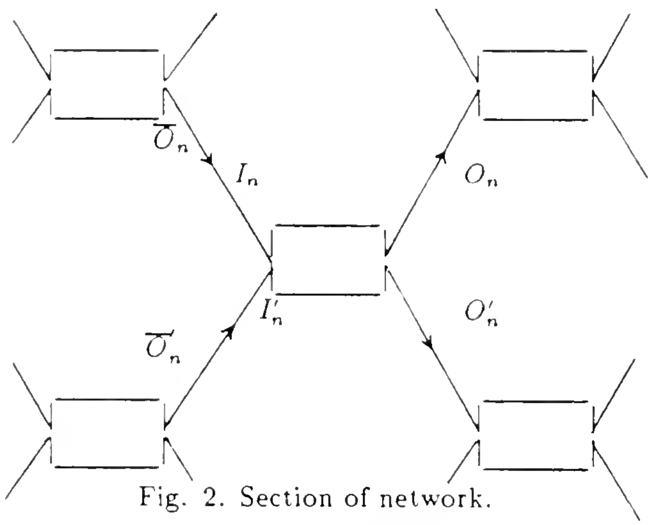

Fig. 2. Section of network.

ρ_0 and ρ_1 of $\{\overline{O}_n\}$ determine the $\{a_n\}$ transition probabilities, all that are required to specify a homogeneous Markov chain. Hence, if we have sufficient information to obtain the corresponding parameters, say ρ_0^* and ρ_1^*, of the output $\{O_n\}$, we will have specified the mixed input $\{a_n\}$ of the next switch, i.e. we will have generated the time series transformation from (ρ_0, ρ_1) to (ρ_0^*, ρ_1^*). All we have to do is find the $\{O_n\}$ transition probabilities.

We start by observing that since $a_1 = 0$, $S_0 = 0$ implies $S_1 = 0$, then

(4.2) $$\begin{aligned} P(O_1 = 0 \mid O_0 = 0) &= P(a_1 = 0,\ S_0 = 0 \mid a_0 = 0,\ S_{-1} = 0) \\ &= P(a_1 = 0 \mid a_0 = 0,\ S_{-1} = 0). \end{aligned}$$

Now S_{-1} is a complicated but explicit function of $a_{-1}, a_{-2}, \ldots$ (assuming for definiteness that $S_{-\infty} = 0$) so that for any time series for which $a_0 = 0$ decouples the future from the past, we have

$$P(O_1 = 0 \mid O_0 = 0) = P(a_1 = 0 \mid a_0 = 0) = P_{00}. \tag{4.3}$$

In particular, this holds if the $\{a_n\}$ constitute a Markov chain. From the definition (3.9) and the fact, (3.11), that $\rho_0^*/\rho_0 = (1 - \rho_1^*)/(1 - \rho_1)$ at fixed ρ, it then follows that

$$\begin{aligned} \rho_0^* &= 1 - P_{00} \\ \rho_1^* &= 1 - \frac{1 - \rho_1}{\rho_0}(1 - P_{00}). \end{aligned} \tag{4.4}$$

Since ρ is conserved, we will instead follow the transformation of the single quantity r of (3.14) to $r^* = (\rho_1^* - \rho)/(2 - \rho)$. Reading off P_{00} from Table 1, we find at once

$$r^* = \frac{q - p}{4q}(p + 2qr + pr^2). \tag{4.5}$$

Turning next to the mean queue length, as given by (2.19), the Markov assumption allows us to write $E(a_i \mid \delta_0 = 1) = E(a_i \mid a_0 = 0,\ S_{-1} = 0)$ as

$$E(a_i \mid \delta_0 = 1) = E(a_i \mid a_0 = 0). \tag{4.6}$$

so that (2.19) relies on the direct properties of the sequence $\{a_n\}$ alone:

$$(q - p)E(S) = p^2 + \sum_{i=1}^{\infty}(E(a_i a_0) - 4p^2) + (q - p)\sum_{i=1}^{\infty}(E(a_i \mid a_0 = 0) - 2p). \tag{4.7}$$

The defining Markov property is that

$$P(a_i = j \mid a_0 = k) = (P^i)_{jk}. \tag{4.8}$$

We also need the (non-hermitian) projection onto the space of eigenvalue 1: since $\sum_k P_{jk} f_k = f_j$, $\sum_j 1 \cdot P_{jk} = 1$, this is given by

$$Q_{jk} = (P^\infty)_{jk} = f_j, \tag{4.9}$$

and satisfies

$$PQ = QP = Q, \quad Q^n = Q. \tag{4.10}$$

We can then write $\sum_{i=1}^{\infty}(E(a_i a_0) - 4p^2) + (q-p)\sum_{i=1}^{\infty}(E(a_i|a_0 = 0) - 2p)$
$= \sum_{i=1}^{\infty}(\sum_k k f_k E(a_i|a_0 = k) - 4p^2 + (q-p)E(a_i \mid a_0 = 0) - 2p(q-p))$
$= \sum_{i=0}^{\infty}(\sum_k [(q-p)\delta_{k,0} + k\, f_k]\; E(a_i \mid a_0 = k) - 2p) + 2p - \sum_k k^2 f_k$ as

$$\sum_{i=0}^{\infty}\sum_{jk} j\Big[(q-p)\delta_{k0} + k\; f_k\Big](P^i - Q)_{jk} - 2p^2 \tag{4.11}$$

There remains the computation of $\sum_{i=0}^{\infty}(P^i - Q) = \sum_{i=0}^{\infty} P^i(I-Q)$. If the cubic characteristic polynomial of P is denoted by $\phi(x)$ with $\phi(1) = 0$, then according to the familiar lagrange interpolation formula, $I - Q = I - \frac{1}{\phi'(1)}\frac{\phi(P)}{P-I} = -\frac{1}{\phi(1)}\,(\phi(P) - \phi(1)I - \phi'(1)(P-I))/(P-I)$ $= -(\frac{1}{2}\phi''(1)(P-I) + \frac{1}{6}\phi'''(1)(P-I)^2)/\phi'(1) = (I-P)(P-I+\frac{1}{2}\phi''(1)I)/\phi'(1)$. But $\sum_{i=1}^{\infty} P^i(I-P) = I - P^{\infty} = I - Q$, and $(I-Q)(P-I) = P - I$, and so

$$\sum_{i=1}^{\infty}(P^i - Q) = \frac{1}{\phi'(1)}[P - I + \frac{1}{2}\phi''(1)(I-Q)] \tag{4.12}$$

Since

$$\begin{aligned}\sum_{jk} j[(q-p)\delta_{k0} + k\; f_k](I-Q)_{jk} &= 2p^2\\ \sum_{jk} j[(q-p)\delta_{k_0} + k\; f_k](P-I)_{jk} &= -2p^2(1-r).\end{aligned} \tag{4.13}$$

(4.7) then becomes

$$(q-p)E(S) = -p^2 + \frac{p^2}{\phi'(1)}(\phi''(1) - 2(1-r)). \tag{4.14}$$

It is not hard to see, from Table 1, that

$$\phi(x) = (x-1)(x-r)(x-r^2); \tag{4.15}$$

hence

$$\phi'(1) = (1-r)^2(1+r), \quad \phi''(1) = 2(2+r)(1-r), \tag{4.16}$$

and (4.14) reduces to

$$E(S) = \frac{1+r}{1-r}\;\frac{p^2}{q-p}. \tag{4.17}$$

or in our original notation, to

$$E(S) = \frac{1+r}{1-r}\,\frac{1}{1-\rho}\,\frac{\rho^2}{4}. \tag{4.18}$$

It is instructive to compare the approximation (4.18), via the transformation (4.5), with computer simulation results of Kruskal et al. At the first stage of the network, with independent inputs $\{I_n, I_n'\}$, then $r = 0$, and (4.18) is exact — see e.g. (2.19). After many stages in Markov approximation, $r^* = r$, and (4.5) informs us that $r^2 - 2\frac{q}{p}\,\frac{1}{q-p}r + 1 = 0$, or

$$\text{asymptotic: } r_\infty = \frac{q}{p}\,\frac{1}{q-p} - [(\frac{q}{p}\,\frac{1}{q-p})^2 - 1]^{1/2}. \tag{4.19}$$

The results, in Table 2, show that the Markov approximation yields a lower bound throughout the region investigated (recent simulations have reinforced this conclusion).

	stage 2			stage 3			stage ∞		
ρ	r	$\overline{E}_{4.18}$	$\overline{E}_{sim}$	r	$\overline{E}_{4.18}$	$\overline{E}_{sim}$	r	$\overline{E}_{4.18}$	$\overline{E}_{sim}$
.2	.011	1.02	1.05	.011	1.02	1.07	.011	1.02	1.08
.5	.042	1.08	1.12	.053	1.11	1.17	.083	1.17	1.20
.8	.033	1.07	1.19	.036	1.07	1.26	.075	1.15	1.30

Table 2. Comparison of Markov approximation (4.18) with simulation results, for $\overline{E} = \frac{4(1-\rho)}{\rho^2}\,E(S)$.

5. A Renewal Model

Three-state discrete Markov processes constitute a rather narrow class of time series, acquiring at most six — and in our realization, only two — parameters for their specification. But of course they have the virtue of simplicity, allowing us to use (4.3) and (4.7) without paying further attention to the precise mechanism of the process. There is however a broader class of time series models which allow the same dramatic simplification of (2.19). It is that of renewal processes, which for our purposes can be defined by the existence of a special state $a = 0$ which decouples the past from the future:

$$\begin{aligned} P(\cdots a_2 = \alpha_2, a_1 = \alpha_1 | a_0 = 0, \ a_{-1} = \alpha_{-1}, \ldots) \\ = P(\cdots a_2 = \alpha_2, \ a_1 = \alpha_1 \mid a_0 = 0) \end{aligned} \tag{5.1}$$

This includes Markov processes as a few-parameter subclass, with innumerable classes which are intermediate. For the analysis we are about to perform, further details will in fact be unnecessary.

Suppose we visualize our queuing network from a typical beginning, where the input time series $\{I_n\}$, and consequently $\{a_n\}$, is one of independent entries and hence trivially renewal. It is not hard to show that with equilibrium current ρ, the first stage output is likewise renewal, characterized by the probability [4],

$$h_n = \frac{\binom{2n+2}{n+1}}{n+2}(1-\frac{\rho}{2})^2(\frac{\rho}{2}(1-\frac{\rho}{2}))^n \tag{5.2}$$

of output $O_0 = 0$ being followed by a string of n values $O_i = 1$ before returning to output 0. If the mixed input $\{a_n\}$ to the next, or second, stage is projected onto the class of renewal processes as well, the information that we then need to evaluate (4.7) can be obtained without difficulty from (5.2). The conclusions we draw will be sufficiently compelling that we will then be excused from the task of characterizing the mixed input in sufficient detail to determine the second stage output, with which the passage through stages can be continued.

To obtain the second stage mean queue length with the present model, we will need the quantities $E(a_i \mid a_0 = 0)$, $E(a_i a_0)$, which we will see in a moment require only

$$\begin{aligned} E(I_i \mid I_0 = 0) &= P(I_i = 1 \mid I_0 = 0) \\ E(I_i I_0) &= P(I_i = 1 \mid I_0 = 1)\rho \ . \end{aligned} \tag{5.3}$$

By copying the argument of (3.8, 3.9, 3.11), it is easy to show that

$$
\begin{aligned}
&\text{if} \quad P(I_i = 0 \mid I_0 = 0) \equiv P_i \\
(5.4) \qquad &\text{then } P(I_i = 1 \mid I_0 = 0) = 1 - P_i \ , \quad P(I_i = 0 \mid I_0 = 1) = \frac{1-\rho}{\rho}(1 - P_i) \\
&P(I_i = 1 \mid I_0 = 1) = \frac{2\rho - 1}{\rho} + \frac{1-\rho}{\rho} P_i \ .
\end{aligned}
$$

There remains only P_i. But by the standard argument which expresses the probability of a return to the origin in termsof the probability of a first return, we know that (see [4])

$$
\begin{aligned}
&\text{if } H(\lambda) = \sum_0^\infty \lambda^n h_n \\
(5.5) \qquad &\text{then } \sum_1^\infty \lambda^i P_i = \lambda H(\lambda)/(1 - \lambda\ H(\lambda)) \ .
\end{aligned}
$$

In particular, for h_n of (5.2), it is easy to see (h_n was in fact obtained originally from $H(\lambda)$) that

$$
(5.6) \qquad \lambda^2 p^2 H(\lambda)^2 + (2\lambda pq - 1)\ H(\lambda) + q^2 = 0 \ ,
$$

where the branch satisfying $H(1) = 1$ is to be chosen.

Now returning to $E(a_i \mid a_0 = 0)$ and $E(a_i a_0)$, we have, according to (3.5), the associated probability generating function

$$
\begin{aligned}
F(x, y) &\equiv \sum_{\alpha, \alpha'} P(a_i = \alpha',\ a_0 = \alpha)\, x^{\alpha'} y^{\alpha} \\
(5.7) \qquad &= \frac{1}{16}(1 + G(x) + G(y) + G_i(x, y))^2 \\
&\text{where } G(x) = 1 + (x - 1)\rho \\
&G_i(x, y) = \sum_{\nu', \nu} P(I_i = \nu',\ I_0 = \nu) x^{\nu'} y^{\nu}
\end{aligned}
$$

Since

$$
\begin{aligned}
&E(a_i \mid a_0 = 0) f_0 = F'_x(1, 0) \\
(5.8) \qquad &E(a_i a_0) = F''_{xy}(1, 1) \ ,
\end{aligned}
$$

it follows that

$$
\begin{aligned}
E(a_i \mid a_0 = 0) f_0 &= \frac{1}{8}(G'(1) + G'_{i,x}(1, 0))\ (1 + G(1) + G(0) + G_i(1, 0)) \\
&= \frac{1}{4}(G'(1) + E(I_i \mid I_0 = 0)\ (1 - \rho))\ (1 + G(0)) \\
(5.9) \qquad &+ \frac{1}{8}(G'(1) + G'_{i,x}(1, 1))\ (G'(1) + G'_{i,y}(1, 1)) \\
&= \frac{1}{2} E(I_i I_0) + \frac{1}{2} G'(1)^2 \ .
\end{aligned}
$$

Inserting (5.4) into (5.9), via (5.3), the relevant combination entering into (4.7) becomes

$$(5.10)\qquad E(a_i a_0) + (q-p)\ E(a_i \mid a_0 = 0) - 2p = \frac{p}{2q}(q-p)(P_i - (q-p))\ .$$

Hence

$$\sum_1^\infty \left(E(a_i a_0) + (q-p)E(a_i \mid a_0 = 0) - 2p\right)\lambda^i = \frac{p}{2q}(q-p)\left(\frac{H(\lambda)}{1-\lambda H(\lambda)} - \frac{q-p}{1-\lambda}\right)\ ,$$

or using (5.6),

$$\begin{aligned} &= \frac{p}{2q}(q-p)\left(\frac{q^2 - \lambda p^2 H(\lambda)}{1-\lambda} - \frac{q-p}{1-\lambda}\right) = \frac{p^3}{2q}(q-p)\frac{1-\lambda H(\lambda)}{1-\lambda} \\ &= \frac{p^3}{2q}(q-p)\frac{H(\lambda)}{q^2 - \lambda p^2 H(\lambda)}\ . \end{aligned}$$

We conclude ontaking $\lambda = 1$ that

$$(5.11)\qquad \sum_1^\infty \left(E(a_i a_0) + (q-p)\ E(a_i \mid a_0 = 0) - 2p\right) = \frac{p^3}{2q}\ ,$$

or according to (4.7)

$$(5.12)\qquad \begin{aligned} E(S) &= \frac{p^2}{q-p}\left(1 + \frac{p}{2q}\right) \\ &= \frac{\rho^2}{4(1-\rho)}\left(1 + \frac{1}{2}\frac{\rho}{2-\rho}\right)\ . \end{aligned}$$

stage 2

ρ	$\overline{E}_{5.12}$	$\overline{E}_{sim}$
0.2	1.05	1.05
0.5	1.17	1.12
0.8	1.33	1.19

Table 3. Comparison of renewal approximation (5.12) with simulation results, for $\overline{E} = \frac{4(1-\rho)}{\rho^2} E(S)$.

The comparison in Table 3 shows that this approximation not only yields an upper bound to $E(S)$ for the full set of currents ρ, but already exceeds the asymptotic result at high ρ. Thus, expansion of the model space provides a considerable overcompensation for the error produced in the Markov model.

By copying the argument of (3.8, 3.9, 3.11), it is easy to show that

$$\text{if} \quad P(I_1 = 0 \mid I_0 = 0) \equiv P_1$$

(5.4)
$$\text{then } P(I_1 = 1 \mid I_0 = 0) = 1 - P_1 \ , \quad P(I_1 = 0 \mid I_0 = 1) = \frac{1-\rho}{\rho}(1 - P_1)$$
$$P(I_1 = 1 \mid I_0 = 1) = \frac{2\rho - 1}{\rho} + \frac{1-\rho}{\rho} P_1 \ .$$

There remains only P_1. But by the standard argument which expresses the probability of a return to the origin in termsof the probability of a first return, we know that (see [4])

(5.5)
$$\text{if } H(\lambda) = \sum_0^\infty \lambda^n h_n$$
$$\text{then } \sum_1^\infty \lambda^i P_i = \lambda H(\lambda)/(1 - \lambda\, H(\lambda)) \ .$$

In particular, for h_n of (5.20) it is easy to see (h_n was in fact obtained originally from $H(\lambda)$) that

(5.6)
$$\lambda^2 p^2 H(\lambda)^2 + (2\lambda pq - 1)\, H(\lambda) + q^2 = 0 \ ,$$

where the branch satisfying $H(1) = 1$ is to be chosen.

Now returning to $E(a_1 \mid a_0 = 0)$ and $E(a_1 a_0)$, we have, according to (3.5), the associated probability generating function

(5.7)
$$F(x, y) \equiv \sum_{\alpha, \alpha'} P(a_1 = \alpha', \ a_0 = \alpha)\, x^{\alpha'} y^{\alpha}$$
$$= \frac{1}{16}(1 + G(x) + G(y) + G_1(x, y))^2$$
$$\text{where } G(x) = 1 + (x - 1)\rho$$
$$G_1(x, y) = \sum_{\nu', \nu} P(I_1 = \nu', \ I_0 = \nu) x^{\nu'} y^{\nu}$$

Since

(5.8)
$$E(a_1 \mid a_0 = 0) f_0 = F'_x(1, 0)$$
$$E(a_1 a_0) = F''_{xy}(1, 1) \ ,$$

it follows that

(5.9)
$$E(a_1 \mid a_0 = 0) f_0 = \frac{1}{8}(G'(1) + G'_{1,x}(1, 0))\ (1 + G(1) + G(0) + G_1(1, 0))$$
$$= \frac{1}{4}(G'(1) + E(I_1 \mid I_0 = 0)\ (1 - \rho))\ (1 + G(0))$$
$$+ \frac{1}{8}(G'(1) + G'_{1,x}(1, 1))\ (G'(1) + G'_{1,y}(1, 1))$$
$$= \frac{1}{2} E(I_1 I_0) + \frac{1}{2} G'(1)^2 \ .$$

Inserting (5.4) into (5.9), via (5.3), the relevant combination entering into (4.7) becomes

$$E(a_i a_0) + (q-p)\ E(a_i \mid a_0 = 0) - 2p = \frac{p}{2q}(q-p)(P_i - (q-p))\ . \tag{5.10}$$

Hence

$$\sum_1^{\infty} (E(a_i a_0) + (q-p)E(a_i \mid a_0 = 0) - 2p)\lambda^i = \frac{p}{2q}(q-p)\left(\frac{H(\lambda)}{1-\lambda H(\lambda)} - \frac{q-p}{1-\lambda}\right)\ ,$$

or using (5.6),

$$\begin{aligned} &= \frac{p}{2q}(q-p)\left(\frac{q^2 - \lambda p^2 H(\lambda)}{1-\lambda} - \frac{q-p}{1-\lambda}\right) = \frac{p^3}{2q}(q-p)\frac{1-\lambda H(\lambda)}{1-\lambda} \\ &= \frac{p^3}{2q}(q-p)\frac{H(\lambda)}{q^2 - \lambda p^2 H(\lambda)}\ . \end{aligned}$$

We conclude ontaking $\lambda = 1$ that

$$\sum_1^{\infty} (E(a_i a_0) + (q-p)\ E(a_i \mid a_0 = 0) - 2p) = \frac{p^3}{2q}\ , \tag{5.11}$$

or according to (4.7)

$$\begin{aligned} E(S) &= \frac{p^2}{q-p}\left(1 + \frac{p}{2q}\right) \\ &= \frac{\rho^2}{4(1-\rho)}\left(1 + \frac{1}{2}\frac{\rho}{2-\rho}\right)\ . \end{aligned} \tag{5.12}$$

stage 2

ρ	$\overline{E}_{5.12}$	$\overline{E}_{\text{sim}}$
0.2	1.05	1.05
0.5	1.17	1.12
0.8	1.33	1.19

Table 3. Comparison of renewal approximation (5.12) with simulation results, for $\overline{E} = \frac{4(1-\rho)}{\rho^2} E(S)$.

The comparison in Table 3 shows that this approximation not only yields an upper bound to $E(S)$ for the full set of currents ρ, but already exceeds the asymptotic result at high ρ. Thus, expansion of the model space provides a considerable overcompensation for the error produced in the Markov model.

6. Discussion

We have treated the problem of passage through a discrete-time clock-regulated queuing network by modeling the input time series $\{a_n\}$ to each queue. Using expected queue length as criterion, the Markov model and renewal model produced small but nontrivial errors on opposite sides, indicating that neither model captures the true flavor of the time series involved. There are at least three options. The first, perhaps tedious but certainly not hard to carry out, is to systematically generalize the Markov model by allowing sucessively longer memory. The second is that of modeling the time series differently, e.g. for output and input, hoping to cancel the under- and over-estimates. However, a compelling rationale for doing so, in the sense of leading to a systematic improvement method, is less easy to formulate. A third approach comes to grips with the specific nature of the networks we have been considering, observing that the innocuous process of mixing output to produce input makes a significant change in the time series. This suggests that it is the simpler two-state output process that should be modeled — indeed this is exactly renewal at the first stage — but raises a host of technical details which are now being studied.

References

[1] Gottlieb, A. An Overview of the N.Y.U. Ultracomputer Project, in: Experimental Parallel Computing Architectures, Dongarra, J. J. Ed., pp. 25-95, 1987.

[2] Hildebrand, F. B., Methods of Applied Mathematics, Prentice Hall, p. 451, 1954.

[3] Kruskal, C. P., Snir, M. and Weiss, A., The Distribution of Waiting Times in Clocked Multistage Interconnection Networks, Procleedings of the 1986 International Conference on Parallel Processing, pp.12-19, 1986.

[4] Percus, O. E., Percus, J. K., Elementary Properties of Clock-regulated Queues, N.Y.U. Ultracomputer Note # 138,1988;
Some Results Concerning Clock-Regulated Queues, Comp. Archit. News, Vol. 16, # 4, pp. 138-145, 1988.

[5] Whittle, P., Systems in Stochastic Equilibrium, Part III, Network Models, John Wiley, 1986.